NVIDIA DGX A100

THE UNIVERSAL SYSTEM FOR AI IN THE PUBLIC SECTOR

The NVIDIA DGX™ A100 offers unmatched compute power and versatility for public sector AI initiatives, enabling every agency to get a faster start in AI, distilling actionable insights from the largest models and datasets, and delivering the performance and efficiency needed to complete the mission. Whether for humanitarian response, national defense, or cybersecurity, DGX A100 offers a purpose-built platform for accelerating the complex computational workloads that power AI development, helping your organization transform into an AI-enabled agency.

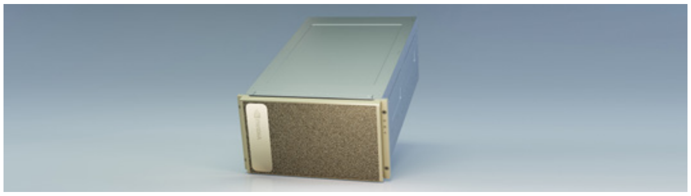
Public Sector Use Cases

- Accelerated processing of large-scale, high-resolution imagery for disaster relief and geospatial intelligence
- Dramatically improving the throughput of analyzing large audio datasets
- Using conversational AI to automate customer assistance
- Analyzing combined multimodal data for pattern detection and behavior prediction
- High-performance data analytics for logistics optimization and platform sustainment
- Streaming analytics for cyber defense

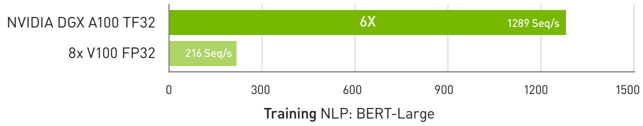
Driving Mission Success with Maximum Efficiency

One of the largest challenges faced by many public sector teams is extracting intelligence, value, and actionable information from large streams or repositories of unstructured data, especially audio and video. Whether trying to analyze full-motion video in a short amount of time, automatically find and categorize figures in patent applications, track packages, or discover if documents contain personally identifiable information (PII) that must be protected, the DGX A100 has the compute power to rapidly train and deploy the most powerful AI models. The DGX A100 offers outstanding performance across the AI development workflow, providing flexible infrastructure that can train complex models at night and become a mirco-services platform for inference by day.

With its high-performance memory and AI compute capacity, DGX A100 is also the optimal platform for data analytics. In-memory databases and graph analytics run orders of magnitude faster on the DGX, whether you're analyzing package logistics by zip code or exploring cyber data for threat hunting.

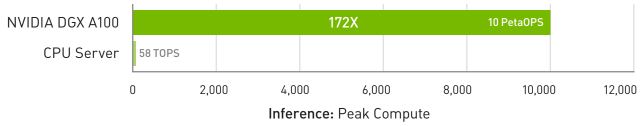


DGX A100 Delivers 6 Times The Training Performance



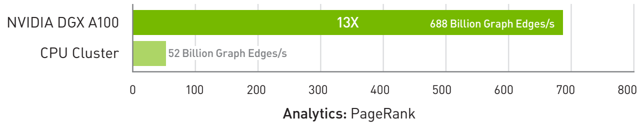
BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 | V100: DGX-1 with 8x V100 using FP32 precision | DGX A100: DGX A100 with 8x A100 using TF32 precision

DGX A100 Delivers 172 Times The Inference Performance



CPU Server: 2x Intel Platinum 8280 using INT8 | DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

DGX A100 Delivers 13 Times The Data Analytics Performance



3000x CPU Servers vs. 4x DGX A100 | Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

Fastest Time to Solution for Public Sector

NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs, which deliver unmatched acceleration, and is fully optimized for NVIDIA CUDA-X™ software and the end-to-end NVIDIA data center solution stack. NVIDIA A100 GPUs bring a new precision, Tensor Float 32 (TF32), which works just like FP32 but provides 20X higher floating operations per second (FLOPS) for AI compared to the previous generation. Best of all, no code changes are required to achieve the speedup. And when using NVIDIA's automatic mixed precision with FP16, A100 offers an additional 2X boost to performance with just one additional line of code.

The A100 GPU has a class-leading 1.6 terabytes per second (TB/s) of memory bandwidth, a greater than 70 percent increase over the last generation. It also has significantly more on-chip memory, including a 40 megabyte (MB) level 2 cache that's nearly 7X larger than the previous generation, maximizing compute performance. DGX A100 also debuts the third generation of NVIDIA® NVLink®, which doubles the GPU-to-GPU direct bandwidth to 600 gigabytes per second (GB/s), almost 10X higher than PCIe Gen 4, and the second generation of NVIDIA NVSwitch™ that is 2X faster than the last generation. This unprecedented power delivers the fastest time to solution, allowing users to tackle challenges that weren't possible or practical before, from searching large audio or video collections to accelerated cyber network mapping.

Unmatched Data Center Scalability with Mellanox

With the fastest input/output (I/O) architecture of any DGX system, NVIDIA DGX A100 is the foundational building block for large AI clusters like NVIDIA DGX SuperPOD™, the enterprise blueprint for scalable AI infrastructure. DGX A100 features eight single-port NVIDIA Mellanox® ConnectX-6 VPI HDR InfiniBand adapters for clustering and one dual-port ConnectX-6 VPI Ethernet adapter for storage and networking, all capable of 200 gigabits per second (Gb/s). The combination of massive GPU-accelerated compute with state-of-the-art networking hardware and software optimizations means DGX A100 can scale to hundreds or thousands of nodes to meet the biggest challenges, such as conversational AI and large-scale image classification.

Proven Infrastructure Solutions Built with Trusted Data Center Leaders

In combination with leading storage and networking technology providers, a portfolio of infrastructure solutions are available that incorporate the best of the NVIDIA DGX POD™ reference architecture. Delivered as fully integrated, ready-to-deploy offerings through the NVIDIA Partner Network (NPN), these solutions simplify and accelerate data center AI deployments.

DGXperts

NVIDIA DGX A100 is more than a server. It's a complete hardware and software platform built upon the knowledge gained from the world's largest DGX proving ground—NVIDIA DGX SATURNV—and backed by thousands of DGXperts at NVIDIA. DGXperts are AI-fluent practitioners who offer prescriptive guidance and design expertise to help fast-track AI transformation. DGXperts help ensure that critical applications get up and running quickly, and stay running smoothly, for dramatically improved time to insights.

SYSTEM SPECIFICATIONS

GPUs	**8x NVIDIA A100 Tensor Core GPUs**
GPU Memory	**320 GB total**
Performance	**5 petaFLOPS AI** **10 petaOPS INT8**
NVIDIA NVSwitches	**6**
System Power Usage	**6.5kW max**
CPU	**Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)**
System Memory	**1TB**
Networking	**8x Single-Port Mellanox ConnectX-6 VPI** **200Gb/s HDR InfiniBand** **1x Dual-Port Mellanox ConnectX-6 VPI** **10/25/50/100/200Gb/s Ethernet**
Storage	**OS: 2x 1.92TB M.2 NVME drives** **Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives**
Software	**Ubuntu Linux OS**
System Weight	**271 lbs (123 kgs)**
Packaged System Weight	**315 lbs (143kgs)**
System Dimensions	**Height: 10.4 in (264.0 mm)** **Width: 19.0 in (482.3 mm) MAX** **Length: 35.3 in (897.1 mm) MAX**
Operating Temperature Range	**5ºC to 30ºC (41ºF to 86ºF)**

To learn more about NVIDIA DGX A100, visit **www.nvidia.com/dgx-a100**
To learn more about our top uses in the public sector, visit **www.nvidia.com/public-sector**

ABC Partner

NVIDIA